深度学习概述

常虹 山世光

摘要 深度学习是机器学习领域的一个新的研究方向，其核心思想在于模拟人脑的层级抽象结构，通过无监督的方式从大规模数据（例如图像、声音和文本）中学习特征。近年来，深度学习在计算机视觉、语音识别等研究领域取得的巨大成功使得研究者们对其寄予更多的关注。本文从深度学习的概念、发展历程、模型、训练方法以及应用等几个方面对其进行概述，并对深度学习的未来发展做出展望。

关键词 深度学习；神经网络；无监督学习；深度置信网络；自动编码器

1 引言

《麻省理工学院技术评论（*Technology Review*）》将深度学习（Deep Learning）列为 2013 年度 10 大技术突破之一。《纽约时报》的头版把深度学习称为一种革命性的人工智能新技术。斯坦福大学教授、著名的机器学习专家吴恩达（Andrew Ng）认为深度学习可以让机器更好地理解人的意图，在未来的 30 到 40 年，深度学习技术有望帮助我们创造出对环境有洞察和学习能力的机器。

那么，何谓深度学习？它何以如此强大？

深度学习是相对于浅层学习而言的。传统的机器学习方法，诸如支持向量机（support vector machines, SVMs）, boosting 等，都是浅层学习方法。在机器学习领域，所谓深度指在一个流向图（flow graph）中的输入到输出的最长路径的长度。例如，SVM 的深度为 2，其中第一层对应其核输出或者特征空间，第二层对应其线性混合的分类输出。传统的前馈神经网络的深度等于其层次的数目。本希奥（Bengio）[4]对深度学习的研究表明，每个函数都有其固定的最小深度，即在运算次序上尽可能并行后的运行次数。函数的深度与所选择的运算有很大的关系。 哈斯塔德（Hastad）等人[9]证明，如果一个函数可以由 k 层网络模型紧致地表示（即通过较少的计算单元），那么用 k-1 层网络模型表示则需要指数倍的计算单元。深层结构可以用少于函数变量和训练数据的计算单元紧致地表示高度变化的函数，这是大多数现有的浅层机器学习方法不可比拟的。所以，本希奥[2]等人认为，增加网络结构的深度从统计学的效率来看是非常重要的。多层函数结构可以增强模型的表达能力并不是一个最近的发现，较早的工作包括引文[29][10]。近来，乌特戈夫（Utgoff）和斯特拉库齐（Stracuzzi）[34]预见较深层的结构在认知方面具有更好的前景，本希奥和勒坤（LeCun）[2]则分析了深层结构的表达能力及其在人工智能和机器学习领域可能的应用。

除了更强大的函数表达能力和更好的泛化能力，深度学习的结果比较自然地体现了底层特征到高层特征的演变。例如，深度模型可以表示“图像块或像素点→边缘→部件→物体”的学习过程，而这个过程与生物的视觉感知系统十分契合。同时，深度学习利用大数据来学习特征，比传统的人工构造特征的方法更能够刻画数据的丰富内在信息，从而最终提升分类或预测的准确性。例如，深度学习系统能够通过扫描无数张猫的图片“认识”猫。从这个意义上说，深度学习也可称为无监督特征学习（unsupervised feature learning）。

深度学习推动图像识别、语音识别等方面的研究取得了突破性的进展，开启了“大数据+复杂模型”的时代。深度学习的胜利应归功于：深度模型结构、高效的学习方法、大数据

的支持以及日新月异的计算能力。

2　深度学习的发展历程

目前，深度模型中可堆叠的学习结构主要是多层神经网络。二十世纪六十年代神经网络第一次兴起，感知机（Perceptron）是其代表性的模型。二十世纪八十年代，第二代神经网络利用反向传播（back propagation, BP）方法学习网络参数，在一定程度上提高了神经网络的性能。但是，BP 算法需要标注数据，可扩展性不好，容易陷入局部极小。之后，瓦普尼克（Vapnik）和他的同事们提出了一种特殊的感知机模型—支持向量机。在这个模型中，每个训练样本产生一维描述测试样本与该训练样本相似程度的特征，模型训练的结果可以找到最佳的特征子集及权重。支持向量机等统计模型在理论分析和应用中都获得了巨大的成功，其优越的性能博得了研究者们的青睐，神经网络的发展进入暂时的沉寂。

神经网络中最为常用的是基于梯度的训练方法，它在训练浅层神经网络（1 或 2 个隐藏层）时具有很好的效果。但本希奥等人通过实验指出，基于梯度的训练方法并不适于多层神经网络。梯度信息在神经网络的较高层中对参数的更新具有很好的指引性。但经过向后传播后，它并不能有效地指导较低层的参数变化到比较理想的区域，这使得整个神经网络很容易陷入到局部最小值中。在很多问题上，往往较多层的神经网络的训练结果反而差于较少层的网络。

2006 年以来，辛顿（Hinton）等一批研究者成功地改变了多层神经网络研究进退唯谷的局面[12][12] [3] [30]，使得深度学习迅速获得了广泛的关注，迎来了机器学习的新浪潮[39]。这种成功主要源于无监督的逐层初始化（layer-wise pre-training），即每一层以较低层的表示作为输入，通过无监督的学习方式训练得到较高（隐藏）层的更抽象的表示形式。无监督学习得到的深层网络参数具有较好的初始值，然后通过自顶向下的监督学习调整网络参数和中间层的特征，避免了反向传播方法的梯度扩散，使其最终和特定的识别任务相关联。

3　深度学习的模型与训练方法

深度学习模型在网络结构上与传统的神经网络相似，都是分层的网络结构。但是，深度学习采用了与神经网络很不同的训练机制。

3.1 主要的深度学习模型

–　卷积神经网络

卷积神经网络（Convolutional neural networks，CNNs）[24]是一种监督学习下的深度模型，最早受视觉系统结构[15]的启发而提出。其基本思想是在前层网络的不同位置共享特征映射的权重，利用空间相对关系减少参数数目以提高训练性能。

卷积神经网络以其局部权值共享的特殊结构和对平移、比例缩放、倾斜等形变的高度不变性，在语音识别

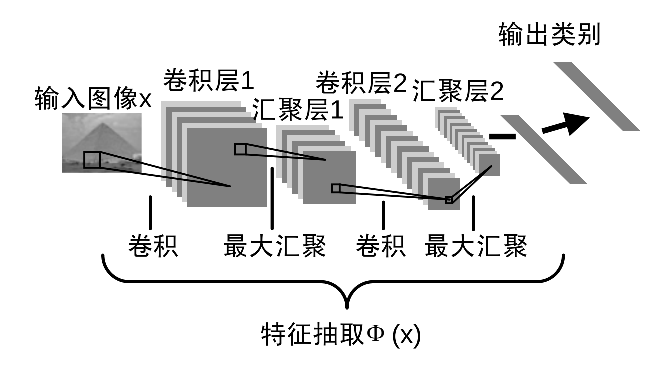


图2.　卷积神经网络抽取图像特征

和图像处理方面显示出独特的优越性。在图像处理方面，输入图像和卷积网络的拓扑结构能很好地吻合，特征提取和模式分类可以同时进行，使神经网络结构变得更简单，适应性更强。因此，基于卷积神经网络的视觉系统至今具有领先的性能，例如，勒坤等人[24][25]改进了卷积神经网络，并在手写字符识别上取得了不错的效果。辛顿等研究者利用一个包含 7 个隐藏层（不包括汇聚层）的卷积神经网络赢得了 ImageNet 竞赛。

杰瑞特（Jarrett）等人[16]发现，即使是单层的未经训练的卷积神经网络，仍可在识别问题上取得很好的效果，不差于，甚至偶尔优于充分训练的卷积神经网络。萨克瑟（Saxe）等人[32]证明，这个现象是由卷积网络结构内在的特性引起的，卷积网络即使未经训练仍具有很好的频率选择性与平移不变性。

– 受限玻尔兹曼机和深度置信网络

受限玻耳兹曼机（Restricted Boltzmann Machine，RBM）如图 2（左）所示，它是玻尔兹曼机的一种变型，即去掉原始的玻尔兹曼机中可见结点之间及隐藏结点之间的连接。受限玻耳兹曼机是一种基于能量的模型，其中二进制神经元的概率值可以通过激励的向上传播和向下传播获得，使用对比离散度（contrastive divergence）方法很大程度上提高了模型的训练效率。

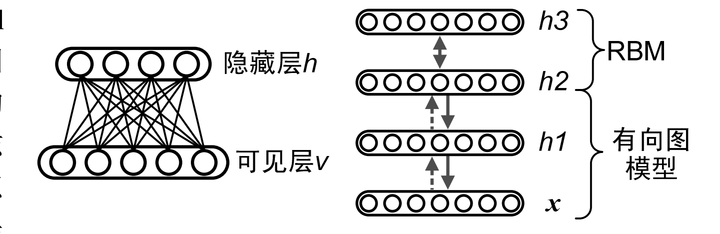


图3. 受限玻尔兹曼机（左）和深度置信网络（右）

受限玻耳兹曼机提供了无监督学习单层网络的方法，如果把隐藏层的层数增加，即得到深度玻尔兹曼机；如果在靠近可见层的部分采用贝叶斯网络（即有向图模型，这里依然限定层中节点之间没有连接），而在最远离可见层的部分使用受限玻耳兹曼机，即得到深度置信网络(Deep Belief Networks, DBNs) [13][3][30]，如图 2（右）所示。深度置信网络可以看作是一种产生式模型，图中实线箭头表示数据产生的过程，虚线箭头表示多层特征提取的过程（或识别过程）。

深度置信网络在模型训练方面充分体现了深度学习的思想：通过自底向上的逐层无监督学习进行初始化，然后通过自顶向下的监督学习微调（fine tune）模型参数。本希奥等[3]提出了针对深度置信网络的逐层贪婪训练方法，即由下到上逐层单独训练受限玻耳兹曼机，其中高层受限玻耳兹曼机训练时的输入由下层训练好的受限玻耳兹曼机传递。虽然深度置信网络已经有许多成功的应用，但是其计算代价高且可扩展性不强。为了使其能够应用于现实的问题中（例如，处理实际尺寸的图像），研究者们提出深度置信网络的若干变形方式，如卷积深度置信网络[26]。

– 自动编码器

另一类深度学习模型以自动编码器（autoencoder，AE）[23][11] [5]为单层网络结构，如图 3（左）所示。与深度置信网络的概率图模型不同，自动编码器通常以重构误差作为优化的目标函数，试图直接学习从输入到输出的、参数化的映射函数（或者特征提取函数）。去噪自动编码器（denoising autoencoder, DAE）[35]是自

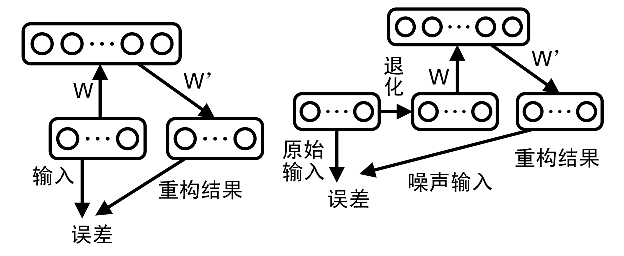


图4. 自动编码器（左)和去噪自动编码器（右）

动编码器的一种随机扩展，它的目标是从有噪声的输入数据中重构原始输入，从而实现更鲁棒的特征学习。

多个自动编码器可以组成叠加自动编码器（stacked autoencoder），并利用深度学习的思想进行训练。为了解决网络退化问题，研究者们通过引入先验提出了稀疏自动编码器[30]、Contractive AE（收缩自动编码器）[31]等变种。

3.2 大规模训练方法

如前文所提到的，深度模型的训练过程包括：（1）通过无监督学习对每一层网络进行初始化，并将其训练结果作为高一层网络的输入；（2）通过监督学习微调整个网络的参数。其中，第一步的初始化是深度学习能取得出色效果的重要因素。深度学习中的优化主要是基于随机梯度下降（stochastic gradient decent, SGD）方法，其参数更新仅在单个训练样本或者一小部分训练样本上进行。虽然随机梯度下降与传统方法更适用于大规模训练数据，但是这种顺序优化的思想增加了并行化的困难，成为该方法在时间效率上面临的最大瓶颈。

深度学习的成功大多建立在“大数据+复杂模型”基础之上，例如谷歌（Google）的科学家们用 1.6 万个处理器构建的深度神经网络能够从 1000 万幅无标注视频帧中学会识别猫的面孔[22]。因此，为了取得更高的性能和效率，发掘更复杂的高层次特征，我们必须提高深度学习方法的可扩展性。目前，训练大规模深度模型（超过 1B[1]参数）主要依赖大量的 CPU 核以及类似云计算的方法。以杰夫.迪恩（Jeff Dean）和吴恩达为首的研究者为了训练具有 1B 参数的大规模深度模型，采用了稀疏、局部感受区域、汇聚（pooling）和局部对比正则化（local contrast normalization）方法，以及模型并行化和异步随机梯度下降（asynchronous SGD）方法[22]。最近，科茨（Coates）等[16]提出基于现成商品高性能计算（Commodity Off-The-Shelf High Performance Computing, COTS HPC）技术的深度学习系统，该系统由无限带宽互联的 GPU 服务器群组成，训练 1B 参数仅需 3 台机器，而且能够扩展到更大的网络规模。

4 深度学习的成功应用

关于深度学习，最令人瞩目的当属其在计算机视觉、语音识别等领域的成功应用。

计算机视觉

在计算机视觉领域，深度学习最初成功的应用是在数据降维[3]、手写数字识别等问题中。近年来，深度学习在更广泛的计算机视觉和模式识别问题中，例如图像识别[26][18]、图像去噪和修复[37]、运动建模[33]、动作识别[17][20][21]、物体跟踪[1] [36]、视觉建模[38] 、场景分析[8]等，展现出了有效性。一个案例是 2012 年，多伦多大学辛顿教授等采用深度卷积神经网络在 ImageNet 图像识别竞赛中将错误率从 26％降低到 15％[19]。

语音识别

2011 年以来，微软研究人员通过与辛顿合作，首先将受限玻耳兹曼机和深度置信网络引入到语音识别的声学模型训练中，在大词汇量语音识别系统中获得巨大成功，使得语音识别的错误率相对降低了约 30%，是语音识别领域十多年来最大的突破性进展[14]。在国际上，IBM、谷歌等公司先后开展了基于深度学习的语音识别研究，并且进展飞快。在国内，百度、

[1] 10^9

科大讯飞、中科院自动化所等公司或研究单位，也开始了深度学习在语音识别上的研究。

其他更多领域

深度学习在自然语言处理方面也具有巨大的潜力[7]，尽管目前的研究还没有取得像语音识别那样的突破性成果。最近，辛顿领导的研究团队基于深度学习方法从大量分子中找到可能成为药物的分子，这项成果由此获得了默克（Merck）公司赞助的一项大奖。事实上，涉及到大数据智能分析和预测的领域都可能找到深度学习的用武之地，这样的领域包括（但不局限于）：互联网行为分析、文本分析、市场监测、自动控制（如无人驾驶汽车）等等。

5 总结与展望

深度学习模拟人脑神经系统构建深层神经网络模型，通过无监督的方式从大量数据中学习层级特征，在计算机视觉、语音识别等领域取得了巨大的成功。可以说，深度学习让我们向真正的智能机时代迈进了一步。但是，深度学习不是一项万能的技术，它能解决的只是构建智能机器所面临的巨大挑战中的一部分。正如纽约大学教授盖瑞.马库斯（Gary Marcus）[28]所言：“辛顿已建立了一个很好的梯子，但这个梯子并不一定能带你到月球。”

基于已有的工作和思考，我们对于深度学习尚未解决的问题和未来的研究方向的看法概括如下：

1. 深度学习尚缺少统计学习理论的有力支持，模型的可表示性、可学习性以及可并行计算性等基础理论问题有待于深入研究。
2. 即使是庞大复杂的深度神经网络，距离模拟真实人脑还差得非常远。我们无法完全掌握人类大脑的工作原理，但是深度学习的成功使得研究者们更加关注脑神经科学的研究，相关的研究项目如雨后春笋般涌现，“大神经科学时代”（Era of Big Neuroscience）已经到来。
3. 深度模型对动态数据建模的成效非常有限，其描述时间序列数据动态特性的能力有待研究。
4. 深度学习在模型训练、观测和解释方面需要进一步的工作，例如中间结果的控制、多层同时训练的方法、深度生成式模型更好的采样方法、模型的解释方法等。
5. 传统浅层学习方法的深度扩展是个值得关注的问题。在大数据时代，深度学习可能并非唯一的选择。一些传统的机器学习方法如何借鉴深度学习的思想解决大数据智能分析问题，也值得研究。
6. 深度模型在特征共享的层面实现了多任务学习的机制，即多个任务之间共享或部分共享较低层的特征表示，而不同的任务对应的高层特征表示各不相同。我们期望未来有更高效的多任务深度学习方法和成功的应用出现。

参考文献：

[1] L. Bazzani, N. Freitas, H. Larochelle, V. Murino, J. Ting. Learning attentional policies for tracking and recognition in video with deep networks. In Proceedings of the 28th International Conference on Machine Learning (ICML), 2011.

[2] Y. Bengio, Y. LeCun. Scaling learning algorithms towards AI. Large Scale Kernel Machines. MIT Press. 2007.

[3] Y. Bengio, P. Lamblin, D. Popovici and H. Larochelle, Greedy Layer-Wise Training of Deep Networks, In Advances in Neural Information Processing Systems (NIPS), pp. 153-160, MIT Press, 2007.

[4] Y. Bengio. Learning deep architectures for AI. Foundations and Trends in Machine Learning, 2(1):1–127, Jan. 2009.

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives. arXiv: 1206.5538v2 [CS.LG], Oct. 2012.

[6] A. Coates, B. Huval, T. Wang, D. Wu, A. Ng, B. Catanzaro. Deep learning with COTS HPC systems. In Proceedings of the 30th International Conference on Machine Learning (ICML), 2013.

[7] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, Pavel Kuksa. Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research, Vol.12, pp. 2493-2537, 2011.

[8] Clement Farabet, Camille Couprie, Laurent Najman and Yann LeCun. Learning Hierarchical Features for Scene Labeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013.

[9] J. Hastad, M. Goldmann. On the power of small-depth threshold circuits. Computational Complexity, 1, 113-129. 1991.

[10] G. Hinton. Connectionist learning procedures. Artificial Intelligence. Vol. 40, 185-234. 1989.

[11] Hinton, G. E. and Zemel, R. S. (1994). Autoencoders, minimum description length, and helmholtz free energy. In Advances in Neural Information Processing Systems (NIPS), 1993.

[12] G. Hinton, R. Salakhutdinov. Reducing the dimensionality of data with neural networks. Science, 313, 504-507, 2006.

[13] G. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. Neural Computation 18:1527-1554, 2006.

[14] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury. Deep Neural Networks for Acoustic Modeling in Speech Recognition. IEEE Signal Processing Magazine. November 2012.

[15] D. Hubel, T. Wiesel. Receptive fields, binocular interaction, and functional architecture in the cats' visual cortex. Journal of Physiology (London), 160, 106-154. 1962.

[16] K. Jarrett, K. Kavukcuoglu, M. Ranzato, Y. LeCun. What is the Best Multi-Stage Architecture for Object Recognition? In Proceedings of International Conference on Computer Vision (ICCV), 2009.

[17] S. Ji, W. Xu, M. Yang, K. Yu. 3D convolutional neural networks for human action recognition. In Proceedings of the 27th International Conference on Machine Learning (ICML), 2010.

[18] K. Kavukcuoglu, P. Sermanet, Y. Boureau, K. Gregor, M. Mathieu, Y. LeCun. Learning Convolutional Feature Hierachies for Visual Recognition. In Advances of Neural Information Processing System (NIPS), 2010.

[19] A. Krizhevsky, I Sutskever, G.Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In Advances in Neural Information Processing Systems (NIPS), 2012.

[20] Q. Le, A. Karpenko, J. Ngiam, A. Ng. ICA with reconstruction cost for efficient overcomplete feature learning. In Advances of Neural Information Processing Systems (NIPS), 2011.

[21] Q. Le, W. Zou, S. Yeung, A. Ng. Learning hierarchical invariant spatio-temporal features for action-recognition with independent subspace analysis. In Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[22] Q. V. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean, A. Ng. Building high-level features using large scale unsupervised learning. In Proceedings of the 29th International Conference on Machine Learning (ICML), 2012.

[23] LeCun, Y. (1987). Mod`eles connexionistes de l'apprentissage. Ph.D. thesis, Universit´e de Paris VI.

[24] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, L. Jackel. Backpropagation

applied to handwritten zip code recognition. Neural Computation, 1(4), 541–551. 1989.

[25] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner. Gradient based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324. 1998.

[26] H. Lee, R. Grosse, R. Ranganath, A. Ng. Convolutional Deep Belief Networks for scalable unsupervised learning of hierarchical representations. In Proceedings of the 26th International Conference on Machine Learning (ICML), 2009.

[27] H. Lee, R. Grosse, R. Ranganath, A. Ng. Convolutional Deep Belief Networks for scalable unsupervised learning of hierarchical representations. In Proceedings of the 26th International Conference on Machine Learning (ICML), 2009.

[28] Marcus, G. (2012). Is "Deep Learning" a Revolution in Artificial Intelligence? The New Yorker. November 25, 2012.

[29] J. McClelland, D. Rumelhart, the PDP Research Group. Parallel distributed processing: explorations in the microstructure of cognition, vol.2. Cambridge: MIT Press. 1986.

[30] M. Ranzato, C. Poultney, S. Chopra and Y. LeCun. Efficient Learning of Sparse Representations with an Energy-Based Model, In Advances in Neural Information Processing Systems (NIPS), MIT Press, 2007.

[31] Rifai, S., Vincent, P., Muller, X., Glorot, X., and Bengio, Y. (2011a). Contractive auto-encoders: Explicit invariance during feature extraction. In Proceedings of the 28th International Conference on Machine Learning (ICML), 2011

[32] A. Saxe, P. Koh, Z. Chen, M. Bhand, B. Suresh, A. Ng. On Random Weights and Unsupervised Feature Learning, In Proceedings of the 28th International Conference on Machine Learning (ICML), 2011.

[33] G. Taylor, G. Hinton, S. Roweis. Modeling human motion using binary latent variables. In Advances of Neural Information Processing Systems (NIPS), 2007.

[34] P. Utgoff, D. Stracuzzi. Many-layered learning. Neural Computation, 14, 2497-2539. 2002.

[35] Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In Proceedings of the 27th International Conference on Machine Learning (ICML), 2008.

[36] Naiyan Wang, Dit-Yan Yeung. Learning a Deep Compact Image Representation for Visual Tracking . To Appear in Proceedings of Twenty-Seventh Annual Conference on Neural Information Processing Systems (NIPS) , Lake Tahoe, Nevada, USA, 5-10 December 2013.

[37] J. Xie, L. Xu, E. Chen. Image denoising and inpainting with deep neural networks. In Advances in Neural Information Processing Systems (NIPS), 12.

[38] W. Zou, S. Zhu, A. Ng, K. Yu. Deep learning of invariant features via simulated fixations in video. In Advances in Neural Information Processing Systems (NIPS), 2012.

[39] 余凯, 深度学习——机器学习的新浪潮，《程序员》，2012.02.

作者简介：

常　虹： 中国科学院计算技术研究所副研究员，changhong@ict.ac.cn

山世光： 中国科学院计算技术研究所研究员，智能信息处理重点实验室常务副主任，sgshan@ict.ac.cn